Large-Context Models for Large-Scale Machine Translation

John DeNero
Dissertation Talk
Statistical Language Systems are Working
Statistical Language Systems are Working

Translation for *donde esta la universidad*:
Spanish » English

*donde esta la universidad* - where is the university
Statistical Language Systems are Working

How?

Data!

Google spent $5.6 billion on infrastructure in the last 3 years.¹

¹ Google.com annual report of capital expenditure, “the majority of which was related to IT infrastructure investments.”
Statistical Language Systems are Working

How?

Translation for **donde esta la universidad**: Spanish » English

**donde esta la universidad** - where is the university

Google spent $5.6 billion on infrastructure in the last 3 years

---

1 Google.com annual report of capital expenditure, “the majority of which was related to IT infrastructure investments.”
Statistical Language Systems are Working

How?

Google spent $5.6 billion on infrastructure in the last 3 years.

---

1 Google.com annual report of capital expenditure, “the majority of which was related to IT infrastructure investments.”
Statistical Language Systems are Working

Google spent $5.6 billion on infrastructure in the last 3 years.¹

How?


People use [Google Translate] hundreds of millions of times a week.” ²

¹ Google.com annual report of capital expenditure, “the majority of which was related to IT infrastructure investments.”
The Many Use(r)s of Machine Translation

<table>
<thead>
<tr>
<th>Assimilation</th>
<th>Dissemination</th>
<th>Communication</th>
</tr>
</thead>
</table>
# The Many Use(r)s of Machine Translation

<table>
<thead>
<tr>
<th>Assimilation</th>
<th>Dissemination</th>
<th>Communication</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Document translation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Broadcast monitoring</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Intelligence gathering</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The Many Use(r)s of Machine Translation

<table>
<thead>
<tr>
<th>Assimilation</th>
<th>Dissemination</th>
<th>Communication</th>
</tr>
</thead>
</table>
| • Document translation
| • Broadcast monitoring
| • Intelligence gathering |
| Most Internet users can’t read the English Web |

73%
The Many Use(r)s of Machine Translation

**Assimilation**
- Document translation
- Broadcast monitoring
- Intelligence gathering

**Dissemination**

**Communication**

Most Internet users can’t read the English Web

73%
# The Many Use(r)s of Machine Translation

<table>
<thead>
<tr>
<th>Assimilation</th>
<th>Dissemination</th>
<th>Communication</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Document translation</td>
<td>- tripadvisor&lt;sup&gt;®&lt;/sup&gt; &lt;br&gt; <em>get the truth. then go.</em>&lt;br&gt; <em>machine translated</em></td>
<td>Most Internet users can’t read the English Web</td>
</tr>
<tr>
<td>- Broadcast monitoring</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Intelligence gathering</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The Many Use(r)s of Machine Translation

Assimilation

- Document translation
- Broadcast monitoring
- Intelligence gathering

Most Internet users can’t read the English Web

Dissemination

Communication

73%

Most Internet users can’t read the English Web
The Many Use(r)s of Machine Translation

Assimilation

- Document translation
- Broadcast monitoring
- Intelligence gathering

Most Internet users can’t read the English Web

Dissemination

tripadvisor
get the truth. then go.™
machine translated

Communication

John I’m in Berkeley
The Many Use(r)s of Machine Translation

Assimilation
- Document translation
- Broadcast monitoring
- Intelligence gathering

Dissemination
- TripAdvisor
  - Get the truth. Then go.
  - Machine translated

Communication
- John
  - I’m in Berkeley
- Juan
  - Estoy en Berkeley

Most Internet users can’t read the English Web
The Many Use(r)s of Machine Translation

Assimilation

- Document translation
- Broadcast monitoring
- Intelligence gathering

Dissemination

- 73% Most Internet users can’t read the English Web

Communication

- Emergency room triage
- Military deployments
- Multilingual education
- 9-1-1 Response
- Commerce with tourists

Example:
- John: I’m in Berkeley
- Juan: Estoy en Berkeley
The Many Use(r)s of Machine Translation

Assimilation
- Document translation
- Broadcast monitoring
- Intelligence gathering

Dissemination
- TripAdvisor
- Translated

Communication
- Emergency room triage
- Military deployments
- Multilingual education
- 9-1-1 Response
- Commerce with tourists

Most Internet users can’t read the English Web

John: I’m in Berkeley
Juan: Estoy en Berkeley

婆婆 [Grandma-in-law]
Data-Driven Machine Translation

Target language corpus gives examples of well-formed sentences

I will get to it later  See you later  He will do it

Parallel corpus gives translation examples

I will do it gladly
Yo lo haré de muy buen grado

You will see later  Después lo veras

Machine translation system:
Data-Driven Machine Translation

Target language corpus gives examples of well-formed sentences
- I will get to it later
- See you later
- He will do it

Parallel corpus gives translation examples
- I will do it gladly
- Yo lo haré de muy buen grado
- You will see later
- Después lo verás

Machine translation system:
Model of translation
Data-Driven Machine Translation

Target language corpus gives examples of well-formed sentences

I will get to it later
See you later
He will do it

Parallel corpus gives translation examples

I will do it gladly
Yo lo haré de muy buen grado
You will see later
Después lo verás

Machine translation system:

Source language
Yo lo haré después

Model of translation

Target language
I will do it later
Stitching Together Fragments

Parallel corpus gives translation examples

YO LO HARÉ DE MUY BUEN GRADO

Después lo verás

Model of translation

Machine translation system:

Yo lo haré de muy buen grado

Después lo verás

I will do it gladly

You will see later

Yo lo haré después

Model of translation

I will do it later
Stitching Together Fragments

Parallel corpus gives translation examples

```
I will do it gladly
```

```
You will see later
```

```
Yo lo haré de muy buen grado
```

```
Después lo verás
```

Machine translation system:

```
Yo lo haré después
```

```
I will do it later
```

---

Diagram showing the process of stitching together fragments with a model of translation.
Stitching Together Fragments

Parallel corpus gives translation examples

Model of translation

Machine translation system:
Stitching Together Fragments

Parallel corpus gives translation examples

I will do it gladly
Yo lo haré de muy buen grado.

You will see later
Después lo verás.

Machine translation system:

Yo lo haré

Model of translation

I will do it later
An Example Syntax-Based Translation

Arabic source sentence:

ورفض الباذ الادلاء بآي تصريحات فور وصوله إلى المقاطعة
An Example Syntax-Based Translation

Arabic source sentence:
ورفض الباز الادلاء بائي تصريحات فور وصوله الى المقاطعة

Reference translation from a human translator:
Al-baz declined to make any statements upon his arrival in the province
An Example Syntax-Based Translation

Arabic source sentence:

ورفض الباز الأدلاء بِإِيَّاِ تَصِيرِيحَاتِ فُوَر وَصُولِهِ إِلَىِ الْمِقَاطِعَة

Reference translation from a human translator:

Al-baz declined to make any statements upon his arrival in the province
Arabic source sentence:
ورفض الباذ الادلاء بآية تصريحات فور وصوله الى المقاطعة

Reference translation from a human translator:
Al-baz declined to make any statements upon his arrival in the province

Features:

<table>
<thead>
<tr>
<th>feature</th>
<th>weight</th>
<th>value</th>
<th>product</th>
</tr>
</thead>
<tbody>
<tr>
<td>derivation-size</td>
<td>0.41</td>
<td>8</td>
<td>3.30</td>
</tr>
<tr>
<td>glue-rule</td>
<td>3.89</td>
<td>2</td>
<td>7.78</td>
</tr>
<tr>
<td>green</td>
<td>-0.08</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>gt_prob</td>
<td>0.40</td>
<td>36.18</td>
<td>14.43</td>
</tr>
<tr>
<td>identity</td>
<td>-9.97</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>is_lexicalized</td>
<td>-0.65</td>
<td>6</td>
<td>-3.91</td>
</tr>
<tr>
<td>lex_pef</td>
<td>1.02</td>
<td>5.47</td>
<td>5.60</td>
</tr>
<tr>
<td>lex_pfe</td>
<td>0.31</td>
<td>4.44</td>
<td>1.39</td>
</tr>
<tr>
<td>lm1</td>
<td>1</td>
<td>22.76</td>
<td>22.76</td>
</tr>
<tr>
<td>lm1-unk</td>
<td>30.08</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>lm2</td>
<td>0.74</td>
<td>26.66</td>
<td>19.79</td>
</tr>
<tr>
<td>lm2-unk</td>
<td>-39.18</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>missingWord</td>
<td>-1.29</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>model1inv</td>
<td>1.02</td>
<td>10.60</td>
<td>10.81</td>
</tr>
<tr>
<td>model1nrm</td>
<td>1.35</td>
<td>11.29</td>
<td>15.22</td>
</tr>
<tr>
<td>nonmonotone</td>
<td>4.17</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>olive</td>
<td>1.95</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>psm1n</td>
<td>0.50</td>
<td>24.65</td>
<td>12.30</td>
</tr>
<tr>
<td>text-length</td>
<td>-3.87</td>
<td>15</td>
<td>-58.05</td>
</tr>
<tr>
<td>trivial_cond_prob</td>
<td>0.41</td>
<td>3.34</td>
<td>1.38</td>
</tr>
<tr>
<td>unk-rule</td>
<td>19.28</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>reported totalcost</td>
<td>52.82</td>
<td>0</td>
<td>52.82</td>
</tr>
</tbody>
</table>
The Steps in a Modern Translation System

- Learn a model
- Apply the model
- Choose a translation
The Steps in a Modern Translation System

Learn a model

Apply the model

Choose a translation

United Nations Proceedings

will do it ADV

VP

lo haré ADV
The Steps in a Modern Translation System

Learn a model

Apply the model

Choose a translation

Yo lo haré después

Later do it I will
The Steps in a Modern Translation System

- Choose a translation
- Learn a model
- Apply the model

---

United Nations Proceedings

---

Yo lo haré después

---

Novel Sentence

---

Later do it I will
I will later do it
That I’ll do later
Later that I’ll do
... I will do it later
...
The Steps in a Modern Translation System

1. Choose a translation
2. Learn a model
3. Apply the model
4. Choose a translation

**Novel Sentence:**

United Nations Proceedings

**VP:**

- will do it: ADV
- lo haré: ADV

**Sentence:**

- Later do it: I will
- I will later do it
- That I’ll do later
- Later that I’ll do...
- I will do it later...
The Steps in a Modern Translation System

1. Choose a translation
2. Learn a model
3. Apply the model
4. Choose a translation

United Nations
Proceedings

Yo lo haré después

Novel Sentence

Learn a model

Apply the model

Choose a translation

Later do it
I will later do it
That I’ll do later
Later that I’ll do
... I will do it later
...
The Steps in a Modern Translation System

- Learn a model
- Apply the model
- Choose a translation
The Alignment Problem in Translation

Thank you, I will do it gladly.

Gracias,
lo haré de muy buen grado.
The Alignment Problem in Translation

Thank you, I will do it gladly.

Gracias, lo haré de muy buen grado.
The Alignment Problem in Translation

Thank you, I will do it gladly.

Gracias, lo haré de muy buen grado.
The Alignment Problem in Translation

Thank you, I will do it gladly.

Thank you
Gracias

Thank you
muy buen

Thank you, I will do it gladly.

Gracias
lo haré de muy buen grado.
The Alignment Problem in Translation

Thank you, I will do it gladly.

About the task:
- A lot can be inferred from lexical statistics
- Correct alignments are not one-to-one
- Some cases are tricky, even for people

Gracias
lo haré de muy buen grado.
The Alignment Problem in Translation

Thank you, I will do it gladly.

Gracias, lo haré de muy buen grado.

About the task:
- A lot can be inferred from lexical statistics
- Correct alignments are not one-to-one
- Some cases are tricky, even for people
The Alignment Problem in Translation

Thank you, I will do it gladly.

About the task:
- A lot can be inferred from lexical statistics
- Correct alignments are not one-to-one
- Some cases are tricky, even for people

About solutions:
- Word-to-word links
- Learning driven by conditional word distributions
Thank you, I will do it gladly.

Gracias, lo haré de muy buen grado.

**About the task:**
- A lot can be inferred from lexical statistics
- Correct alignments are not one-to-one
- Some cases are tricky, even for people

**About solutions:**
- Word-to-word links
- Learning driven by conditional word distributions
Thank you, I will do it gladly.

Gracias, lo haré de muy buen grado.
**Large-Context Alignment Challenges**

**Goal:** Model multi-word structures during alignment

$$P(\text{gracias}|\text{you}) \quad P(\text{gracias, Thank you})$$

- **Challenge**
  - Jointly infer phrase boundaries and alignments
  - Boundaries depend on both languages

Thank you, I will do it gladly.

Gracias, lo haré de muy buen grado.
Large-Context Alignment Challenges

**Goal:** Model multi-word structures during alignment

\[
P(\text{gracias} \mid \text{you}) \quad P(\text{gracias}, \text{Thank you}) \quad \phi(\text{lo haré}, \text{I will do it})
\]

**Challenge 1**
- Jointly infer phrase boundaries and alignments
- Boundaries depend on both languages

**Challenge 2**
- Capture context
- Compose phrases
Modeling Phrasal Correspondence

*Paradigm:* Train a generative model that explains observed translations via latent structure
Modeling Phrasal Correspondence

**Paradigm:** Train a generative model that explains observed translations via latent structure

**Process:** Phrase pairs are generated independently
Modeling Phrasal Correspondence

Paradigm: Train a generative model that explains observed translations via latent structure

Process: Phrase pairs are generated independently

Thank you, I will do it gladly

Gracias, lo haré de muy buen grado
Modeling Phrasal Correspondence

Paradigm: Train a generative model that explains observed translations via latent structure

Process: Phrase pairs are generated independently

Thank you, I will do it gladly

Gracias, lo haré de muy buen grado
Modeling Phrasal Correspondence

Paradigm: Train a generative model that explains observed translations via latent structure

Process: Phrase pairs are generated independently

Thank you, I will do it gladly

Gracias, lo haré de muy buen grado
Modeling Phrasal Correspondence

Paradigm: Train a generative model that explains observed translations via latent structure

Process: Phrase pairs are generated independently

Thank you, I will do it gladly

Gracias, lo haré de muy buen grado
Modeling Phrasal Correspondence

**Paradigm:** Train a generative model that explains observed translations via latent structure

**Process:** Phrase pairs are generated independently

Thank you, I will do it gladly

Gracias, lo haré de muy buen grado
Modeling Phrasal Correspondence

Paradigm: Train a generative model that explains observed translations via latent structure

Process: Phrase pairs are generated independently
Modeling Phrasal Correspondence

Paradigm: Train a generative model that explains observed translations via latent structure

Process: Phrase pairs are generated independently

Optimization: Explain all translations with shared parameters
We learn $\theta$, a multinomial distribution over phrase pairs

$$P(A = a) = \theta(\text{Thank you, Gracias}) \cdot \theta(\text{I will do, haré}) \cdot \theta(\text{it, lo}) \cdots$$
Modeling Phrasal Correspondence

We learn $\theta$, a multinomial distribution over phrase pairs

$$P(A = a) = \theta(\text{Thank you, Gracias}) \cdot \theta(\text{I will do, haré}) \cdot \theta(\text{it, lo}) \cdots$$

*Terms omitted: Phrase pair count and phrase permutation*
We learn $\theta$, a multinomial distribution over phrase pairs

$$P(A = a) = \theta(\text{Thank you, Gracias}) \cdot \theta(\text{I will do, haré}) \cdot \theta(\text{it, lo}) \cdots$$

$$\mathcal{L}(\theta) = \prod_{d \in D} \left[ \sum_{a \in A(d)} P(A = a) \right]$$

*Terms omitted: Phrase pair count and phrase permutation*
Modeling Phrasal Correspondence

We learn $\theta$, a multinomial distribution over phrase pairs

$$P(A = a) = \theta(\text{Thank you, Gracias}) \cdot \theta(\text{I will do, haré}) \cdot \theta(\text{it, lo}) \cdots$$

For each sentence pair:

$$\mathcal{L}(\theta) = \prod_{d \in D} \left[ \sum_{a \in A(d)} P(A = a) \right]$$

*Terms omitted: Phrase pair count and phrase permutation*
Modeling Phrasal Correspondence

We learn \( \theta \), a multinomial distribution over phrase pairs

\[
P(A = a) = \theta(\text{Thank you, Gracias}) \cdot \theta(\text{I will do, haré}) \cdot \theta(\text{it, lo}) \cdots
\]

\[
P(A = a) = \prod_{(e, s) \in a} \theta(e, s)
\]

For each sentence pair:

For each alignment:

\[
\mathcal{L}(\theta) = \prod_{d \in D} \left[ \sum_{a \in A(d)} P(A = a) \right]
\]

*Terms omitted: Phrase pair count and phrase permutation*
Modeling Phrasal Correspondence

We learn \( \theta \), a multinomial distribution over phrase pairs

\[
P(A = a) = \prod_{(e,s) \in a} \theta(e, s) \quad * 
\]

For each sentence pair:

For each alignment:

\[
\mathcal{L}(\theta) = \prod_{d \in D} \left[ \sum_{a \in A(d)} P(A = a) \right] 
\]

Maximizing likelihood gives a degenerate solution: huge phrases!

* Terms omitted: Phrase pair count and phrase permutation
Modeling Phrasal Correspondence

We learn \( \theta \), a multinominal distribution over phrase pairs

\[
P(A = a) = \theta(\text{Thank you, Gracias}) \cdot \theta(\text{I will do, haré}) \cdot \theta(\text{it, lo}) \cdots
\]

For each sentence pair:

\[
\mathcal{L}(\theta) = \prod_{d \in D} \left[ \sum_{a \in A(d)} P(A = a) \right]
\]

Maximizing likelihood gives a degenerate solution: huge phrases!

*Terms omitted: Phrase pair count and phrase permutation*
Guiding Phrasal Correspondence Models

\[ \theta \sim \text{DP}(\theta_0, \alpha) \]

**Base distribution:** \( \theta_0 \) \hspace{2cm} \text{Prefers short phrases}

**Dirichlet process:** \( \text{DP}(\cdot, \alpha) \) \hspace{1cm} \text{Non-parametric cache model}

<table>
<thead>
<tr>
<th>English-Spanish phrase pair</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Thank you, Gracias)</td>
<td>1111</td>
</tr>
<tr>
<td>(Thanks, Gracias)</td>
<td>111</td>
</tr>
<tr>
<td>(Thank you, Muchas gracias)</td>
<td>11</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Guiding Phrasal Correspondence Models

\[ \theta \sim \text{DP}(\theta_0, \alpha) \]

**Base distribution:** \[ \theta_0 \] \hspace{1cm} Prefers short phrases

**Dirichlet process:** \[ \text{DP}(\cdot, \alpha) \] \hspace{1cm} Non-parametric cache model

Phrase Pair Cache (c):

<table>
<thead>
<tr>
<th>English-Spanish phrase pair</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>(Thank you, Gracias)</td>
<td></td>
</tr>
<tr>
<td>(Thanks, Gracias)</td>
<td></td>
</tr>
<tr>
<td>(Thank you, Muchas gracias)</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Guiding Phrasal Correspondence Models

\[ \theta \sim \text{DP}(\theta_0, \alpha) \]

**Base distribution:** \[ \theta_0 \]

**Dirichlet process:** \[ \text{DP}(\cdot, \alpha) \]

Prefers short phrases

Non-parametric cache model

Thank you, I will do it gladly.

Gracias

Phrase Pair Cache (c):

<table>
<thead>
<tr>
<th>English-Spanish phrase pair</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>(Thank you, Gracias)</td>
<td>ⅠⅠⅠⅠ</td>
</tr>
<tr>
<td>(Thanks, Gracias)</td>
<td>ⅠⅠ</td>
</tr>
<tr>
<td>(Thank you, Muchas gracias)</td>
<td>ⅠⅠ</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Guiding Phrasal Correspondence Models

\[ \theta \sim \text{DP}(\theta_0, \alpha) \]

**Base distribution:** \( \theta_0 \)

**Dirichlet process:** \( \text{DP}(\cdot, \alpha) \)

Prefers short phrases

Non-parametric cache model

Phrase Pair Cache (c):

<table>
<thead>
<tr>
<th>English-Spanish phrase pair</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Thank you, Gracias)</td>
<td>1111</td>
</tr>
<tr>
<td>(Thanks, Gracias)</td>
<td>1111</td>
</tr>
<tr>
<td>(Thank you, Muchas gracias)</td>
<td>11</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Guiding Phrasal Correspondence Models

\[ \theta \sim \text{DP}(\theta_0, \alpha) \]

**Base distribution:** \[ \theta_0 \]  
Prefers short phrases

**Dirichlet process:** \[ \text{DP}(\cdot, \alpha) \]  
Non-parametric cache model

English-Spanish phrase pair | Count
---|---
(Thank you, Gracias) | 1111
(Thanks, Gracias) | 111
(Thank you, Muchas gracias) | 11
... | ...

\[ P(z|c) = \frac{c(z) + \alpha \cdot \theta_0(z)}{|c| + \alpha} \]
Guiding Phrasal Correspondence Models

\[ \theta \sim \text{DP}(\theta_0, \alpha) \]

**Base distribution:** \[ \theta_0 \]

**Dirichlet process:** \[ \text{DP}(\cdot, \alpha) \]

Prefers short phrases

Non-parametric cache model

\[ P(z|c) = \frac{c(z)}{|c|} + \frac{\alpha \cdot \theta_0(z)}{\alpha} \]

Phrase Pair Cache (c):

<table>
<thead>
<tr>
<th>English-Spanish phrase pair</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Thank you, Gracias)</td>
<td>1111</td>
</tr>
<tr>
<td>(Thanks, Gracias)</td>
<td>111</td>
</tr>
<tr>
<td>(Thank you, Muchas gracias)</td>
<td>11</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Guiding Phrasal Correspondence Models

\[ \theta \sim \text{DP}(\theta_0, \alpha) \]

**Base distribution:** \[ \theta_0 \]

**Dirichlet process:** \[ \text{DP}(\cdot, \alpha) \]

Prefers short phrases

Non-parametric cache model

Phrases Pair Cache (c):

<table>
<thead>
<tr>
<th>English-Spanish phrase pair</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Thank you, Gracias)</td>
<td>1111</td>
</tr>
<tr>
<td>(Thanks, Gracias)</td>
<td>111</td>
</tr>
<tr>
<td>(Thank you, Muchas gracias)</td>
<td>11</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>

\[
P(z | c) = \frac{c(z)}{|c|} + \frac{\alpha \cdot \theta_0(z)}{\alpha}
\]
Guiding Phrasal Correspondence Models

\[ \theta \sim \text{DP}(\theta_0, \alpha) \]

**Base distribution:** \( \theta_0 \)  
Prefers short phrases

**Dirichlet process:** \( \text{DP}(\cdot, \alpha) \)  
Non-parametric cache model

\[
P(z|c) = \frac{c(z)}{|c|} + \frac{\alpha \cdot \theta_0(z)}{\alpha}
\]

Phrase Pair Cache (c):

<table>
<thead>
<tr>
<th>English-Spanish phrase pair</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Thank you, Gracias)</td>
<td>1:1:1:1</td>
</tr>
<tr>
<td>(Thanks, Gracias)</td>
<td>1:1:1</td>
</tr>
<tr>
<td>(Thank you, Muchas gracias)</td>
<td>1:1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Thank you, I will do it gladly.

Gracias, lo haré de muy buen grado.
Guiding Phrasal Correspondence Models

$$\theta \sim DP(\theta_0, \alpha)$$

**Base distribution:** $\theta_0$

**Dirichlet process:** $DP(\cdot, \alpha)$

Prefers short phrases

Non-parametric cache model

![Phrase Pair Cache (c):](image)

Iterative realignment of all the data by sampling

Consistent, efficient estimation
What Happens in Practice

A state-of-the-art word-level alignment

A sampled phrase alignment from our system

Thank you, I shall do so gladly.

Gracias, lo haré de muy buen grado.
Thank you, I shall do so gladly.

A state-of-the-art word-level alignment

A sampled phrase alignment from our system

Gracias, lo haré de muy buen grado.
Performance Results

Translation performance in a phrase-based system (Moses) for Spanish-to-English parliamentary proceedings (Europarl)

- **Word-level baseline**
- **Phrase-level model** [DeNero et al. EMNLP ’08]*

### Translation quality (BLEU)

- **29.8**
- **30.1**

### Millions of phrase pairs

- **4.4**
- **3.1**

Subsequent Work

We described a non-parametric Bayesian prior and a consistent sampling procedure (EMNLP 2008)

- Trevor Cohn and Phil Blunsom. A Bayesian Model of Syntax-Directed Tree to String Grammar Induction, EMNLP 2009.
- Phil Blunsom and Trevor Cohn. Inducing Synchronous Grammars with Slice Sampling, NAACL 2010.
A Model of Composed Phrases

<table>
<thead>
<tr>
<th>In</th>
<th>the</th>
<th>past</th>
<th>two</th>
<th>years</th>
</tr>
</thead>
</table>

过去 [past]
两 [two]
年 [year]
中 [in]
A Model of Composed Phrases

In the past two years

过去 [past]
两 [two]
年 [year]
中 [in]

In the past two years
A Model of Composed Phrases

In the past two years

过去 [past]
两 [two]
年 [year]
中 [in]
A Model of Composed Phrases

过去 [past]
两 [two]
年 [year]
中 [in]

In the past two years
A model can predict the whole analysis above, including minimal links & composed phrase pairs.

过去 [past]
两 [two]
年 [year]
中 [in]

In the past two years
A model can predict the whole analysis above, including minimal links & composed phrase pairs.
A Model of Composed Phrases

Features on word links:
\[ \text{In} \mid \text{中} = 0.8, \quad \text{InDictionary} = 1.0, \quad \ldots \]

Features on phrase pairs:
\[ \text{Count}(\text{the past two}, \text{过去 两}) = 7, \quad \text{Size}(3,2) = 1, \quad \ldots \]

A model can predict the whole analysis above, including minimal links \[ \square \] & composed phrase pairs \[ \square \].
Learning from Supervised Data

Guess: Model Prediction

In the past two years
Learning from Supervised Data

**Guess:** Model Prediction

**Gold:** Human Annotation

In the past two years

- In
- the
- past
- two
- years

过去 [past]

两年 [two]

中 [in]
Learning from Supervised Data

**Guess:** Model Prediction

**Gold:** Human Annotation

In the past two years

过去 [past]
两年 [two]
中 [in]
Learning from Supervised Data

**Guess: Model Prediction**

**Gold: Human Annotation**

In the past two years

In the past two years
Learning from Supervised Data

**Guess:** Model Prediction

In the past two years

**Gold:** Human Annotation

过去 [past]
两年 [two]
中 [in]

In the past two years

**Loss function:** Number of differing rounded rectangles
Learning from Supervised Data

**Guess: Model Prediction**

**Gold: Human Annotation**

*Loss function*: Number of differing rounded rectangles

Online learning (MIRA) adjusts model parameters to prefer the *gold* over the *guess* by a margin of the loss.
Finding the Optimal Correspondence

\[
\arg\ max_{y \in \text{ITG}(x)} \theta \cdot \left[ \phi_{\text{word}}(x, y) + \phi_{\text{phrase}}(x, y) \right]
\]
Finding the Optimal Correspondence

$$\arg \max_{y \in \text{ITG}(x)} \theta \cdot [ \phi_{\text{word}}(x, y) + \phi_{\text{phrase}}(x, y) ]$$

Hierarchical decomposition

<table>
<thead>
<tr>
<th>过去</th>
<th>[past]</th>
</tr>
</thead>
<tbody>
<tr>
<td>两</td>
<td>[two]</td>
</tr>
<tr>
<td>年</td>
<td>[year]</td>
</tr>
<tr>
<td>中</td>
<td>[in]</td>
</tr>
</tbody>
</table>

In the past two years
Finding the Optimal Correspondence

\[
\text{arg max}_{y \in \text{ITG}(x)} \theta \cdot [ \phi_{\text{word}}(x, y) + \phi_{\text{phrase}}(x, y) ]
\]
Finding the Optimal Correspondence

$$\arg \max_{y \in \text{ITG}(x)} \theta \cdot \left[ \phi_{\text{word}}(x, y) + \phi_{\text{phrase}}(x, y) \right]$$
Finding the Optimal Correspondence

\[
\text{arg max}_{y \in \text{ITG}(x)} \theta \cdot \left[ \phi_{\text{word}}(x, y) + \phi_{\text{phrase}}(x, y) \right]
\]
Finding the Optimal Correspondence

$$\arg \max_{y \in \text{ITG}(x)} \theta \cdot [ \phi_{\text{word}}(x, y) + \phi_{\text{phrase}}(x, y) ]$$
Finding the Optimal Correspondence

\[
\arg \max_{y \in ITG(x)} \theta \cdot \left[ \phi_{\text{word}}(x, y) + \phi_{\text{phrase}}(x, y) \right]
\]
Finding the Optimal Correspondence

$$\arg \max_{y \in \text{ITG}(x)} \theta \cdot \left[ \phi_{\text{word}}(x, y) + \phi_{\text{phrase}}(x, y) \right]$$
Finding the Optimal Correspondence

\[
\text{arg max}_{y \in \text{ITG}(x)} \ \theta \cdot \left[ \phi_{\text{word}}(x, y) + \phi_{\text{phrase}}(x, y) \right]
\]
Finding the Optimal Correspondence

$$\arg \max_{y \in \text{ITG}(x)} \theta \cdot [ \phi_{\text{word}}(x, y) + \phi_{\text{phrase}}(x, y) ]$$

Hierarchical decomposition

In the past two years
Finding the Optimal Correspondence

\[
\arg \max_{y \in \text{ITG}(x)} \theta \cdot \left[ \phi_{\text{word}}(x, y) + \phi_{\text{phrase}}(x, y) \right]
\]

Hierarchical decomposition

In the past two years

ITG parser with a state space that tracks peripheral alignments for each region
## Experimental Results

<table>
<thead>
<tr>
<th>Model Type</th>
<th>Alignment Quality</th>
<th>Translation Quality for Chinese-to-English</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unsupervised word model baseline</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Supervised word model [Haghighi, Blitzer, DeNero, and Klein. ACL ’09]*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Composed Phrase Pair Model [DeNero and Klein. In submission]**</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Phrase Pair F1</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>64.1</td>
<td>34.5</td>
</tr>
<tr>
<td>68.4</td>
<td>34.7</td>
</tr>
<tr>
<td>71.6</td>
<td>35.9</td>
</tr>
</tbody>
</table>


The Steps in a Modern Translation System

- Learn a model
- Apply the model
- Choose a translation
The Steps in a Modern Translation System

- Learn a model
- Apply the model
- Choose a translation

- Large data sets provide statistics for larger structures
The Steps in a Modern Translation System

- Learn a model
- Apply the model
- Choose a translation

- Large data sets provide statistics for larger structures
- Non-parametric models scale with the data
The Steps in a Modern Translation System

- Learn a model
- Apply the model
- Choose a translation

- Large data sets provide statistics for larger structures
- Non-parametric models scale with the data
- The more context we incorporate, the better we do
The Steps in a Modern Translation System

Learn a model  Apply the model  Choose a translation
Thank you, I will do it gladly.

Gracias, lo haré de muy buen grado.
Thank you, I will do it gladly.

Gracias, lo haré de muy buen grado.
Extracting Translation Rules

Thank you, I will do it gladly.

Gracias, lo haré de muy buen grado.
Thank you, I will do it gladly.

Gracias, lo haré de muy buen grado.
Thank you, I will do it gladly.

Gracias, lo haré de muy buen grado.
Thank you, I will do it gladly.

Gracias, lo haré de muy buen grado.
Extracting Translation Rules

Thank you, I will do it gladly.

Gracias, lo haré de muy buen grado.

Frequency statistics on these rules guide translation.
Synchronous Context-Free Grammars

<table>
<thead>
<tr>
<th>Grammar</th>
</tr>
</thead>
<tbody>
<tr>
<td>Derivation</td>
</tr>
</tbody>
</table>

Translation:

Source: Mi dormitorio nuevo no es ni grande ni pequeño
Mi dormitorio nuevo no es ni grande ni pequeño

Grammar
Derivation

Translation:

Source: Mi dormitorio nuevo no es ni grande ni pequeño
Synchronous Context-Free Grammars

Grammar

Derivation

Translation:

Source: Mi dormitorio nuevo no es ni grande ni pequeño
Synchronous Context-Free Grammars

Grammar

Derivation

Translation:

**NN**

bedroom

dormitorio

Source: Mi dormitorio nuevo no es ni grande ni pequeño
Mi dormitorio nuevo no es ni grande ni pequeño

Gramática

Derivación

Traducción:
Synchronous Context-Free Grammars

Grammar

Derivation

Translation:

Source: Mi dormitorio nuevo no es ni grande ni pequeño
Synchronous Context-Free Grammars

Grammar

Derivation

Translation: new bedroom big small

Source: Mi dormitorio nuevo no es ni grande ni pequeño
Synchronous Context-Free Grammars

Grammar

Derivation

Translation: My new bedroom

Source: Mi dormitorio nuevo no es ni grande ni pequeño
Synchronous Context-Free Grammars

Grammar

Derivation

Translation: My new bedroom

Source: Mi dormitorio nuevo no es ni grande ni pequeño
Synchronous Context-Free Grammars

Grammar:

Mi dormitorio nuevo no es ni grande ni pequeño

Derivation:

Translation: My new bedroom big small

Source: Mi dormitorio nuevo no es ni grande ni pequeño
Mi dormitorio nuevo no es ni grande ni pequeño

**Grammar**
- **NN** dormitorio
- **JJ** nuevo, grande, pequeño

**Derivation**
- **S** → **NP**
  - **JJ** new
  - **NN** bedroom
  - **JJ** big, small

**Translation:**
- My new bedroom is neither big nor small

**Source:**
- Mi dormitorio nuevo no es ni grande ni pequeño
The Size of the Grammar

A grammar learned from 220 million words of Arabic-to-English example translations:

332,000 rules match a 30-word sentence to be translated

The Size of the Grammar

A grammar learned from 220 million words of Arabic-to-English example translations:

332,000 rules match a 30-word sentence to be translated

The Structure of the Grammar

\[ S \rightarrow NP \text{ no es ni } JJ \text{ ni } JJ \]

Mi dormitorio nuevo no es ni grande ni pequeño
The Structure of the Grammar

\[ S \rightarrow NP \text{ no es ni } JJ \text{ ni } JJ \]

Mi dormitorio nuevo no es ni grande ni pequeño
The Structure of the Grammar

\[ S \rightarrow NP \text{ no es ni } JJ \text{ ni } JJ \]

\[ ADJP \rightarrow \text{ ni } JJ \text{ ni } JJ \]

Mi dormitorio nuevo no es ni grande ni pequeño
The Structure of the Grammar

\[ S \rightarrow NP \text{ no es ni } JJ \text{ ni } JJ \]

\[ ADJP \rightarrow \text{ ni } JJ \text{ ni } JJ \]

\[ VP \rightarrow \text{ es } ADJP \]

Mi dormitorio nuevo no es ni grande ni pequeño
The Structure of the Grammar

\[ S \rightarrow NP \text{ no es ni } JJ \text{ ni } JJ \]

\[ ADJP \rightarrow \text{ ni } JJ \text{ ni } JJ \]

\[ VP \rightarrow \text{ es } ADJP \]

\[ S \rightarrow NP \text{ no } VP \]

Mi dormitorio nuevo no es ni grande ni pequeño
Mi dormitorio nuevo no es ni grande ni pequeño
Coarse-to-Fine Translation

1. Apply a subset of the grammar with only small rules

Mi dormitorio nuevo no es ni grande ni pequeño
Coarse-to-Fine Translation

1. Apply a subset of the grammar with only small rules

Mi dormitorio nuevo no es ni grande ni pequeño
Coarse-to-Fine Translation

1. Apply a subset of the grammar with only small rules
2. Prune away unlikely portions of the search space

Mi dormitorio nuevo no es ni grande ni pequeño
Coarse-to-Fine Translation

1. **Apply a subset of the grammar with only small rules**

2. **Prune away unlikely portions of the search space**

---

Mi dormitorio nuevo no es ni grande ni pequeño

```
S

NP

NN  JJ

VP

ADJP

JJ  JJ
```
Coarse-to-Fine Translation

1. **Apply a subset of the grammar with only small rules**

2. **Prune away unlikely portions of the search space**

**Mi dormitorio nuevo no es ni grande ni pequeño**
Coarse-to-Fine Translation

1. Apply a subset of the grammar with only small rules
2. Prune away unlikely portions of the search space
3. Apply the full translation grammar to the pruned space

Mi dormitorio nuevo no es ni grande ni pequeño
Experimental Results

Minutes required to analyze a 300 sentence test set

<table>
<thead>
<tr>
<th>Method</th>
<th>Time (minutes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>264</td>
</tr>
<tr>
<td>+Optimization</td>
<td>181</td>
</tr>
<tr>
<td>+Transformation</td>
<td>104</td>
</tr>
<tr>
<td>+Coarse-to-Fine</td>
<td>50</td>
</tr>
</tbody>
</table>

5x speed-up with the largest translation grammars in use today (ISI Syntax-Based MT System) [DeNero et al. NAACL ’09]*

The Steps in a Modern Translation System

Learn a model

Apply the model

Choose a translation
The Steps in a Modern Translation System

Learn a model

Apply the model

Choose a translation

- Fully exploiting large data sets requires searching over very large spaces
The Steps in a Modern Translation System

Learn a model

Apply the model

Choose a translation

- Fully exploiting large data sets requires searching over very large spaces

- Coarse-to-fine inference is a powerful technique for doing so
The Steps in a Modern Translation System

- Learn a model
- Apply the model
- Choose a translation
Even the Best Models are Wrong

Yo lo haré después

Model
Even the Best Models are Wrong

Yo lo haré después

Later do it I will
I will later do it
That I’ll do later
Later that I’ll do
...

Model
Even the Best Models are Wrong

Model

Later do it I will
I will later do it
That I’ll do later
Later that I’ll do ...

Total model score for 1000 sentences

Translation Quality (BLEU)

- Samples from output space
- Samples near maximum
- Highest scoring translation

Yo lo haré después

Novel sentence
Even the Best Models are Wrong

Model

Yo lo haré después

Later do it I will
I will later do it
That I’ll do later
Later that I’ll do

...
Even the Best Models are Wrong

Model

Yo lo haré después

Later do it I will
I will later do it
That I’ll do later
Later that I’ll do ...

Translation Quality (BLEU)

+$+$ Samples from output space
+$+$ Samples near maximum
+$+$ Highest scoring translation

Total model score for 1000 sentences
Even the Best Models are Wrong

Model

Yo lo haré después

Later do it I will
I will later do it
That I’ll do later
Later that I’ll do ...

Samples from output space
Samples near maximum
Highest scoring translation

Translation Quality (BLEU)

Total model score for 1000 sentences
Even the Best Models are Wrong

Model

- Later do it I will
- I will later do it
- That I’ll do later
- Later that I’ll do ...

---

Samples from output space

Samples near maximum

Highest scoring translation

Translation Quality (BLEU)

Total model score for 1000 sentences
Consensus by Averaging Over Sentences

Model

Later do it I will
I will later do it
That I’ll do later
Later that I’ll do
...

Yo lo haré después
Consensus by Averaging Over Sentences

Intuition: “Happy families are all alike; every unhappy family is unhappy in its own way.” [Tolstoy. 1877]*

* Leo Tolstoy. Анна Каренина. 1877.
Consensus by Averaging Over Sentences

Intuition: “Happy families are all alike; every unhappy family is unhappy in its own way.” [Tolstoy. 1877]*

Idea: Average over sentences to find the phrases that are alike. [DeNero et al. ACL ’09]**

* Leo Tolstoy. Анна Каренина. 1877.
Consensus by Averaging Over Sentences

Intuition: “Happy families are all alike; every unhappy family is unhappy in its own way.” [Tolstoy. 1877]*

Idea: Average over sentences to find the phrases that are alike. [DeNero et al. ACL ’09]**

“Later” “do” … “do it” “I’ll” “do later”…“do it I will”

* Leo Tolstoy. Анна Каренина. 1877.
Consensus by Averaging Over Sentences

**Intuition:** “Happy families are all alike; every unhappy family is unhappy in its own way.” [Tolstoy. 1877]*

**Idea:** Average over sentences to find the phrases that are alike. [DeNero et al. ACL ’09]**

```
Later” “do” ... “do it” “I’ll” “do later” ...“do it I will”
```

```
1 1 1 1 0 0 1
```

---

* Leo Tolstoy. Анна Каренина. 1877.

Consensus by Averaging Over Sentences

Intuition: “Happy families are all alike; every unhappy family is unhappy in its own way.” [Tolstoy. 1877]*

Idea: Average over sentences to find the phrases that are alike. [DeNero et al. ACL ’09]**

<table>
<thead>
<tr>
<th>“Later”</th>
<th>“do”</th>
<th>…</th>
<th>“do it”</th>
<th>“I’ll”</th>
<th>“do later”</th>
<th>…</th>
<th>“do it I will”</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Leo Tolstoy. Анна Каренина. 1877.
Consensus by Averaging Over Sentences

Intuition: “Happy families are all alike; every unhappy family is unhappy in its own way.” [Tolstoy. 1877]*

Idea: Average over sentences to find the phrases that are alike. [DeNero et al. ACL ’09]**

```
<table>
<thead>
<tr>
<th>“Later”</th>
<th>“do”</th>
<th>“do it”</th>
<th>“I’ll”</th>
<th>“do later”</th>
<th>“do it I will”</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
```

* Leo Tolstoy. Анна Каренина. 1877.

Intuition: “Happy families are all alike; every unhappy family is unhappy in its own way.” [Tolstoy. 1877]*

Idea: Average over sentences to find the phrases that are alike. [DeNero et al. ACL ’09]**

<table>
<thead>
<tr>
<th>“Later”</th>
<th>“do”</th>
<th>...</th>
<th>“do it”</th>
<th>“I’ll”</th>
<th>“do later”</th>
<th>“do it I will”</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.12</td>
<td>1</td>
<td></td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0.10</td>
<td>1</td>
<td></td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.07</td>
<td>1</td>
<td></td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0.07</td>
<td>1</td>
<td></td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

* Leo Tolstoy. Анна Каренина. 1877.
Consensus by Averaging Over Sentences

Intuition: “Happy families are all alike; every unhappy family is unhappy in its own way.” [Tolstoy. 1877]*

Idea: Average over sentences to find the phrases that are alike. [DeNero et al. ACL ’09]**

<table>
<thead>
<tr>
<th>“Later”</th>
<th>“do”</th>
<th>“do it”</th>
<th>“I’ll”</th>
<th>“do later”</th>
<th>“do it I will”</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1 0 0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1 0</td>
<td>0 1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>1 1</td>
<td>0 1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>1 0</td>
<td>0 1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Expected output

* Leo Tolstoy. Анна Каренина. 1877.

Phrase Expectations from Forests
Phrase Expectations from Forests

\[ \text{Yo} \quad \text{lo} \quad \text{haré} \quad \text{después} \]
Phrase Expectations from Forests

I

shall do it

NP

Yo

VP

lo

haré

después
Yo haré después

Yo [NP] I

lo [PRP] it

shall do it [VP] shall do it lo haré

después
Phrase Expectations from Forests
Phrase Expectations from Forests
Phrase Expectations from Forests

"I shall"

"I will"

"do it"
Phrase Expectations from Forests

Yo

I {will, shall} do it

S

NP VP

NP VP

“will do it”

will do PRP

PRP

it

lo

“do it”

Yo

lo

haré

después
Phrase Expectations from Forests

I {will, shall} do it later

I {will, shall} do it

“I will”

will do it

“it later”

“do it”

I

shall do it

NP  \(\downarrow\) I

Yo

lo

haré

después

PRP  \(\uparrow\) it

it

VP  \(\downarrow\) will do PRP haré

VP  \(\uparrow\) shall do it lo haré

S  \(\uparrow\) NP VP

NP VP

S  \(\uparrow\) NP VP

S later

S después

PRP  \(\uparrow\) it

PRP
Single System Translation Results

Translation quality in ISI’s Full-Scale Arabic-to-English Hierarchical Translation System

- Model-Only Baseline
- Consensus from a List [DeNero et al. ACL ’09]*
- Consensus from a Forest [DeNero et al. ACL ’09]*

Translation Quality (BLEU)

- Model-Only Baseline: 52.0
- Consensus from a List: 52.2
- Consensus from a Forest: 53.0

Translating Using Multiple Systems

Yo lo haré después

Model 1
Model 2
Model 3
Translating Using Multiple Systems

Yo lo haré después

Model 1

Model 2

Model 3
Translating Using Multiple Systems

Yo lo haré después

Model 1

Model 2

Model 3

0.96 0.99
0.94 0.92
0.97 0.98
0.54 0.41 0.34 0.12
0.57 0.44 0.30 0.00
Translating Using Multiple Systems

Model 1

Model 2

Model 3

Yo lo haré después

I will do it later
Consensus Modeling FAQ

Q: How do we combine different models?
Consensus Modeling FAQ

Q: How do we combine different models?
A: Train a linear consensus model scoring a derivation $d$:

$$
\sum_{i=1}^{I} \left[ w_i^{(\alpha)} \alpha_i(d) + \sum_{n=1}^{4} w_i^{(n)} v_i^{(n)}(d) \right] + w^{(b)} \cdot b(d) + w^{(\ell)} \cdot \ell(d)
$$
**Consensus Modeling FAQ**

**Q:** How do we combine different models?

**A:** Train a linear consensus model scoring a derivation $d$:

$$
\sum_{i=1}^{I} \left[ w_{i}^{(\alpha)} \alpha_{i}(d) + \sum_{n=1}^{4} w_{i}^{(n)} v_{i}^{(n)}(d) \right] + w^{(b)} \cdot b(d) + w^{(\ell)} \cdot \ell(d)
$$

<table>
<thead>
<tr>
<th>Models</th>
<th>Which model?</th>
<th>Phrase lengths</th>
<th>Expected counts</th>
<th>Model score</th>
<th>Length</th>
</tr>
</thead>
</table>

**Q:** What output sentences are considered?
Q: How do we combine different models?
A: Train a linear consensus model scoring a derivation $d$:

$$
\sum_{i=1}^{I} \left[ w_i^{(\alpha)} \alpha_i(d) + \sum_{n=1}^{4} w_i^{(n)} v_i^{(n)}(d) \right] + w^{(b)} \cdot b(d) + w^{(\ell)} \cdot \ell(d)
$$

Q: What output sentences are considered?
A: The union of output spaces of models:
Multi-System Translation Results

Google’s Full-Scale Research Translation System for Arabic-to-English

Best Single-System Model-Only Baseline
Multi-System Forest-Based Consensus [DeNero et al. NAACL ’10]*

Translation quality (BLEU)

46
45
44
43
42

43.9
45.3

The Steps in a Modern Translation System

1. Learn a model
2. Apply the model
3. Choose a translation
The Steps in a Modern Translation System

Learn a model

Apply the model

Choose a translation

- Statistical models provide distributions over outputs
The Steps in a Modern Translation System

1. Learn a model
2. Apply the model
3. Choose a translation

- Statistical models provide distributions over outputs
- Leveraging those distributions improves performance
The Steps in a Modern Translation System

- Learn a model
- Apply the model
- Choose a translation

- Statistical models provide distributions over outputs
- Leveraging those distributions improves performance
- Compact representations can enable large-scale computation
Summary of Translation Research

Learn a model

Apply the model

Choose a translation
Summary of Translation Research

- Large-context models
- Non-parametric models

Learn a model  Apply the model  Choose a translation

[DeNero et al. EMNLP ’08]
[DeNero & Klein. ACL ’10]
Summary of Translation Research

- Large-context models
- Non-parametric models
- Coarse-to-fine

Learn a model  Apply the model  Choose a translation

[DeNero et al. EMNLP ’08]  [DeNero et al. NAACL ’09]
[DeNero & Klein. ACL ’10]
Summary of Translation Research

- Large-context models
- Non-parametric models
- Coarse-to-fine
- Full distributions
- Compact encodings

Learn a model

[DeNero et al. EMNLP ’08]
[DeNero & Klein. ACL ’10]

Apply the model

[DeNero et al. NAACL ’09]

Choose a translation

[DeNero et al. ACL ’09]
[DeNero et al. NAACL ’10]
Summary of Translation Research

- Large-context models
- Non-parametric models
- Coarse-to-fine
- Full distributions
- Compact encodings

Learn a model

[DeNero et al. EMNLP ’08]
[DeNero & Klein. ACL ’10]

Apply the model

[DeNero et al. NAACL ’09]

Choose a translation

[DeNero et al. ACL ’09]
[DeNero et al. NAACL ’10]

Are we done yet?
Summary of Translation Research

- Large-context models
- Non-parametric models
- Coarse-to-fine
- Full distributions
- Compact encodings

Learn a model

[DeNero et al. EMNLP '08]
[DeNero & Klein. ACL '10]

Apply the model

[DeNero et al. NAACL '09]
[DeNero et al. ACL '09]

Choose a translation

[DeNero et al. NAACL '10]

- Morphology in alignment modeling

Are we done yet?
Summary of Translation Research

- Large-context models
- Non-parametric models
- Coarse-to-fine
- Full distributions
- Compact encodings

Learn a model

- DeNero et al. EMNLP ’08
- DeNero & Klein. ACL ’10

Apply the model

- DeNero et al. NAACL ’09

Choose a translation

- DeNero et al. ACL ’09
- DeNero et al. NAACL ’10

Are we done yet?

- Morphology in alignment modeling
- Unsupervised composed phrase learning
Summary of Translation Research

- Large-context models
- Non-parametric models
- Coarse-to-fine
- Full distributions
- Compact encodings

Learn a model
- [DeNero et al. EMNLP ’08]
- [DeNero & Klein. ACL ’10]

Apply the model
- [DeNero et al. NAACL ’09]

Choose a translation
- [DeNero et al. ACL ’09]
- [DeNero et al. NAACL ’10]

Are we done yet?
- Morphology in alignment modeling
- Unsupervised composed phrase learning
- Adding information to consensus models
Acknowledgements

Thank you!

and many thanks to my excellent coauthors on this work:

 Berkeley: Mohit Bansal, John Blitzer, Alex Bouchard-Côté, Aria Haghighi, Dan Klein, and Adam Pauls

 Information Sciences Institute: David Chiang and Kevin Knight

 Google: Cipriani Chelba, Shankar Kumar, and Franz Och